IMAGE STRUCTURE MODELS OF TEXTURE AND CONTOUR VISIBILITY

Wilson S. Geisler, Thomas Thornton, Donald P. Gallogly & Jeffrey S. Perry
Department of Psychology and Center for Vision and Image Sciences
University of Texas at Austin, Austin TX 78712
geisler@psy.utexas.edu

1. SUMMARY

The perceptual mechanisms underlying texture and contour grouping/segregation play a dominant role in determining the visibility of targets in complex backgrounds. In most quantitative models of texture segregation the image is initially processed by channels selective along certain fundamental stimulus dimensions such as spatial frequency and orientation. These channels generally contain a nonlinearity, such as full-wave rectification, so that they signal the local contrast energy within the bandpass of the channel. Another stage of linear filtering, followed by a simple edge finding or thresholding mechanism, is then applied to the channel outputs to find the texture boundaries or regions. Although these channel-energy models have been successful in predicting texture segregation and discrimination performance for some classes of stimuli, there are large classes of stimuli that are readily segregated by human observers but which cannot be segregated by channel energy. The evidence suggests that more sophisticated models incorporating perceptual organization mechanisms will be required to predict human texture and contour segregation performance. This paper describes new experimental evidence, and a working model which, in principle, can account for a wider range of human segregation and grouping capabilities. The premise of the model is that the visual system typically extracts rich descriptions of local image structure, and that it uses these descriptions for subsequent segregation and grouping. The model contains physiologically-based low level mechanisms for extracting primitives, matching mechanisms for detecting structural similarity, and grouping mechanisms for binding structural parts into wholes. Quantitative predictions of the model for contour segregation performance are presented.

2. INTRODUCTION

"Bottom-up" mechanisms for grouping and segregation are absolutely essential to object detection and recognition. To recognize an object in a typical natural environment, the features of the object must be segregated, at least to some extent, from those of the surrounding objects and surfaces.

In recent years, most models of grouping/segregation have been based upon mechanisms which compare, across space, the contrast energy within spatial-frequency and orientation tuned channels (e.g., for reviews see, Bergen, 1991; Bovik, Clark, & Geisler, 1990; Graham, Beck and Sutter, 1992). While such grouping/segregation mechanisms may exist within the human visual system there are at least two grouping abilities they cannot explain. First, humans are able to segregate image regions based upon differences in the local spatial structure, even when the channel energies are the same (Thornton, et al. 1998; see later). Second, humans are able to detect spatial structure and statistical regularities that vary smoothly over space (i.e., non-stationary image structure). The most well-known example of this is the ability of humans to detect a contour formed by a sequence of line segments (a dashed contour) embedded in a background of randomly oriented line segments (e.g., Sha'ashua & Ullman, 1988; Field, Hayes & Hess, 1993; see later).

The aim of this paper is to demonstrate some of the weaknesses of the *channel energy models* and to demonstrate how some of those weaknesses might be addressed by models incorporating mechanisms which explicitly extract and use local image structure—*image structure models*. We begin by briefly describing a generalized channel energy model and an image structure model.

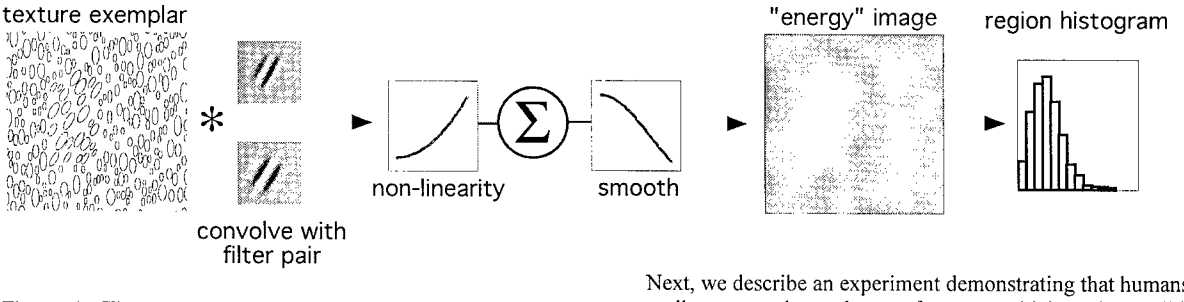

Figure 1. The generalized channel energy model for segregation and grouping. The input image is a complete set of filter pairs each tuned to a different spatial frequency and orientation. One filter pair is shown in the figure. The responses of each filter in the pair are squared and then summed, mimicking the response of a complex cortical cell in area V1. The responses (either with or without smoothing) form an "energy" image. In the most general case, the responses in a region are represented by a histogram of the response magnitudes taken over the region.

Next, we describe an experiment demonstrating that humans can easily segment large classes of textures which are impossible even for an optimal channel energy model. We also show that an image structure model can segment such textures. We then describe parametric measurements of contour detection performance and show that a very simple image structure model is able to account for most aspects of the data.

2.1 Generalized Channel Energy Model

Figure 1 illustrates the generalized channel energy model of segregation and grouping. In this illustration, the input image consists of a target region of diagonal ellipses in a background of vertical ellipses. The input is processed by 30 separate spatial-

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

frequency and orientation tuned channels (6 frequencies x 5 orientations), with spatial frequency bandwidths of 1 octave, and orientation bandwidths of 30 deg. The quadrature pair of receptive fields corresponding to one of the channels is illustrated in the second panel. The channel energy at each pixel location in the image is obtained by summing the square of the responses from the two quadrature components. One can think of the channel energy at a pixel as a response similar to the one that would be produced by a complex cell in primary visual cortex centered on that pixel. Next, the channel energy may be smoothed, or not smoothed, depending on the specific version of the model. As can be seen, the channel energy is greatest in the region of the image where the ellipse orientation is similar to that of the channel. In the generalized energy model, the responses in a region are represented by a response histogram tallied over all the pixel locations in the region. In this illustration, the response range was divided into 15 equal-width regions. In a more conventional energy model, the responses in a region are represented by the sum or average response in the region (i.e., the mean of the histogram). The generalized histogram model extracts some local spatial phase information and hence predicts that humans can discriminate a wider range of textures than predicted by a conventional energy model. Nonetheless, the experiment described later shows that there are many textures humans can segregate, which the ideal generalized channel energy model predicts cannot be segregated.

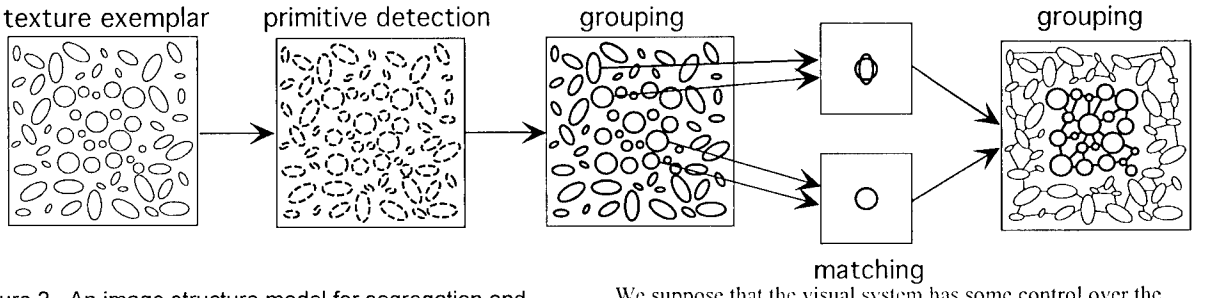

Figure 2. An image structure model for segregation and grouping. The input image is encoded into local primitives (in this case oriented line segments). Initial groups are formed by associative grouping. The initial groups are compared by a matching process to measure their similarities/differences. Higher order groups are obtained by another stage of simple or associative grouping. The processes of matching and grouping may be repeated.

2.2 An Image Structure Model

Figure 2 illustrates an image structure model of segregation and grouping. In this illustration, the input image consists of a target region of circles in a background of randomly oriented ellipses. The input image is encoded as a collection of local primitives (oriented pieces of contour). Different choices for the primitives are possible; in one version, we obtain the primitives by thresholding the responses of model simple cells tuned to different orientations (Geisler & Super, 1996/98). The primitives are represented by the small segments in the second panel. Next, associative grouping of the primitives (described below) is used to obtain initial groups, which, in this case, correspond to the "elements." The initial groups are then compared with one another by a matching process, under some family of transformations (e.g., translation, rotation and scaling). The matching process produces measures of the similarities/differences between the initial groups. In this case, the element shapes within the target and background regions match each other, but do not match across the regions. Finally, the grouping process is applied to the initial groups using the similarities/differences found by the matching process. Grouping on the basis of shape similarity will correctly segment the texture regions. This texture is difficult to segment on the basis of generalized channel energy because of the random sizes, positions and orientations of the elements. We now describe some of the more important components of the model in a bit more detail.

2.3.1 Similarity Measure

To determine the structural relationships between groups, there must be some mechanism for measuring the similarities (or equivalently, differences) of the groups along relevant stimulus dimensions. We assume this is done in parallel across the groups (e.g., across the "elements" in the third panel of Figure 2) by a matching process. In the current version of the model, we suppose that the matching process measures differences between the groups along stimulus dimensions which include form/shape (D_f), position (D_p), orientation (D_θ), size (D_s), symmetry (D_r), and continuation (D_c). The definitions of these *group differences* are described in more detail in Geisler & Super (1996/99). Later we give the definitions for position and continuation. Our working hypothesis is that the potential for binding the two groups together is given by the *total group difference*, which is a linear weighted sum of the group differences:

$$D = w_f D_f + w_p D_p + w_\theta D_\theta + w_s D_s + w_r D_r + w_c D_c \quad (1)$$

We suppose that the visual system has some control over the weights and hence may favor some dimensions under certain situations.

2.3.2 Simple and Associative Grouping

We assume that there are two basic grouping mechanisms which the visual system may use. The first is to bind together groups for which the total group differences are small. We call this *simple grouping*. More formally, if the total grouping difference (D_{ij}) between groups g_i and g_j falls below some criterion (β) then the groups are bound together:

$$if \quad D_{ij} < \beta \quad then \quad g_i \circ g_j \quad (2)$$

where the symbol "$\circ$" represents the operation of binding. In this definition, D_{ij} may represent the weighted sum of grouping differences for all the dimensions except position (proximity). For simple grouping we assume that the weight on the position difference is zero. This assumption is required in order for simple grouping to extend across significant spatial distances.

Associative grouping combines proximity grouping with simple grouping and a transitivity rule. Just as in simple grouping, if the total grouping difference between groups g_i and g_j falls below some criterion then the groups are bound together:

$$if \quad D_{ij} < \beta \quad then \quad g_i \circ g_j \quad (3)$$

In addition, if group i binds to group j and group j binds to group k then groups i and k are bound together:

$$if \quad g_i \circ g_j \,\&\, g_j \circ g_k \quad then \quad g_i \circ g_k \qquad (4)$$

In this definition, D_{ij} represents the total grouping difference, which includes the position/proximity grouping difference.

Our working hypothesis is that the visual system tries a number of values of the binding criterion either simultaneously or sequentially, and then picks values based upon three rules: the *stability rule*, the *performance rule*, and the *recognition rule*. The stability rule is a "bottom up" rule which depends upon the dynamics of group formation. When the binding criterion is varied there will be ranges in the value of the binding criterion where the pattern of grouping is changing rapidly and there will be ranges where the grouping is stable. The stability rule is to pick values from ranges where the grouping is stable (see Estabrook, 1966 for a similar concept of classification). The performance rule is a "top down" rule where the criterion is adjusted or selected based upon improving task performance. The recognition rule is a top down rule which depends upon feedback from subsequent recognition processes. The rule is to pick values of the binding criterion that yield recognized objects/parts when the groups are analyzed further; we suppose that this further recognition analysis is occurring simultaneously while the binding criterion is being varied.

2.3.3 Repeated Grouping

An important aspect of the image structure model is that the grouping processes are carried out repeatedly. First, associative or simple grouping is applied to the detected primitives to find initial groups. After matching, which provides new grouping differences, associative or simple grouping is applied again to obtain higher order groups. Although not indicated in Figure 2, there may be additional repetitions of matching and grouping. It is the repeated applications of matching and grouping which provide the detailed description of image structure.

3. EXPERIMENT 1

Thornton & Geisler (1998) showed that conventional channel energy models predict that certain classes of texture are impossible to segregate when, in fact, humans find them easy to segregate. The classes of textures they considered are similar to those in Figure 2 (see also Victor & Brodie, 1978). The target and background regions consisted of elements that differed in shape, but were randomized in orientation, size, and position.

In the present experiment we tested whether similar results hold for generalized channel energy models, which assume that the human visual system uses the additional information which is contained in channel response histograms computed over texture regions. The generalized channel energy models are more powerful and hence more difficult to reject. To test these models we developed a special procedure for constructing the texture segregation stimuli.

3.1 Methods

The logic of the experiment is quite simple. Construct texture segregation stimuli for which the optimal generalized channel energy model is at chance performance in a forced choice task. If the human observers can perform the segregation task at above chance then the general class of channel energy models is rejected. The difficult part is in constructing the stimuli.

3.1.1 Task

The task was to decide whether a rectangular target region, filled with one texture, was oriented vertically or horizontally within a background region, filled with another texture. The location of the target region was random from trial to trial.

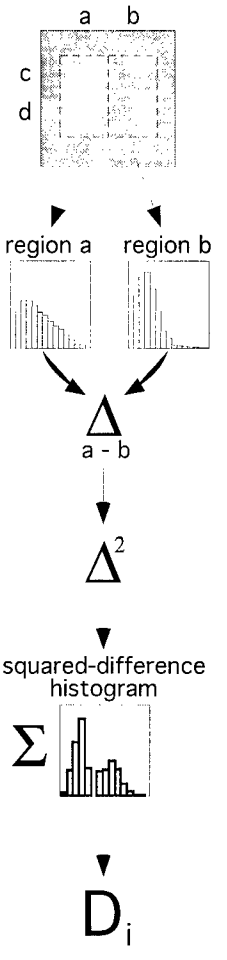

Figure 3. Initial processing steps of the general channel energy model applied to the segregation task of Experiment 1. For each channel-energy "image" a histogram is computed for regions a and b. The histograms are differenced, squared and summed to obtain a measure, D_i, of the histogram difference for each channel.

3.1.2 Stimuli

The exemplars for any given texture stimulus were created by filling a rectangular "target" region (2 by 4 deg) with elements of one shape, and filling the square "background" region (8 by 8 deg) with elements of a another shape. The shapes were always smoothly connected contours. They were generated by summing sinewave components with frequencies in orientation that were harmonics of one cycle per 360 deg. Different random shapes were obtained by randomly selecting the radial amplitudes and phases of the components and then filtering (i.e., multiplying the amplitudes by a transfer function). Exemplar texture stimuli were created by filling each region (target or background) with non-overlapping elements having a single shape, but random size, position, and orientation. For the purposes of creating the stimuli, each element was represented by a virtual circumscribing box. The virtual boxes were randomly placed one at a time within the image, with the restriction that if a new virtual box overlapped an

existing one then a new element was selected. The center of the circumscribing box was allowed to just touch the invisible border defining the regions; this resulted in texture regions with a "natural," irregular appearing boundary. In order to insure that segmentation of target and surround was due solely to specified regional differences in local shape, first order cues of luminance and contrast were balanced across regions and exemplars by keeping pixel density constant.

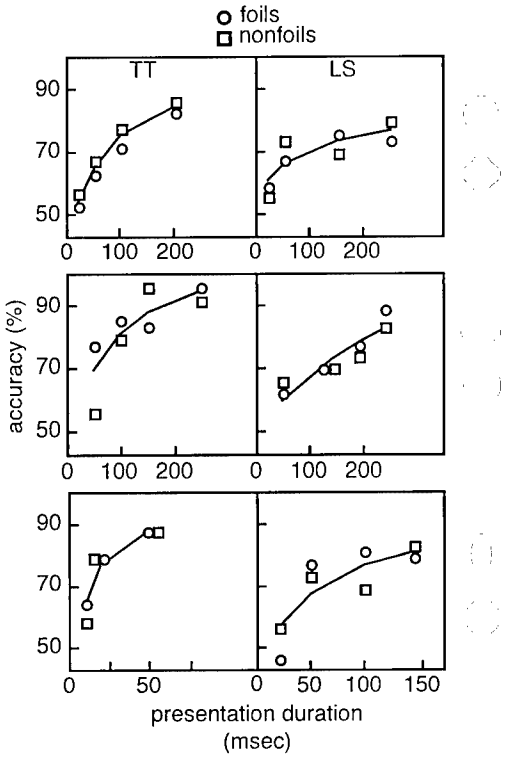

Figure 4. Texture segregation accuracy as a function of stimulus duration for two subjects. The shapes on the right indicate the shapes of elements in the target and background regions. The size, orientation and position of the elements in each region were random. The optimal channel energy model is at chance performance for these stimuli.

From many texture exemplars generated in the above fashion, we selected a subset for which the generalized channel energy model was at chance. To do this we simply collected exemplars that were misclassified (i.e. "foils"). These "foils" were then combined with a number of correctly-classified exemplars to form a stimulus ensemble that by definition yields 50% correct performance. In the experiments reported here, all stimulus ensembles were selected in this manner.

The key step in generating the stimulus ensembles was determining the optimal performance for the general channel energy model in the segregation task. Figure 3 illustrates the initial sequence of processing for the segregation task. On each trial, a set of 30 energy "images" were computed as per the process described in Figure 1 (one image for each of the channels comprising the model's front-end). The first panel in Figure 3 represents one of these energy images. To be conservative, we determined model performance assuming that the target could appear in only one of four locations: *a*, *b*, *c*, and *d* in Figure 3. (The human observers were confronted with greater uncertainty because the targets were randomly positioned within the central region.)

Optimal performance was achieved by applying a maximum likelihood decision rule to the channel responses produced on each trial. First, we computed the sum of the squared difference in the histograms for regions *a* and *b*, for each channel. In Figure 3, D_i represents the value of this quantity for the i^{th} channel. Then, we computed the probability that these 30 values were generated by a vertically oriented target region and by a horizontally oriented target region. The response was which ever orientation was more probable.

The probability densities for the two orientations were assumed to be adequately represented by 30-dimensional multivariate normal density functions (one dimension for each channel) with arbitrary mean vectors and covariance matrices. The mean vector and covariance matrix for vertical targets was estimated from a large number of exemplar vertical stimuli. Similarly, the mean vector and covariance matrix for horizontal targets was estimated from a large number of exemplar horizontal stimuli.

3.1.3 Procedure

Observers made forced choice target orientation decisions ("vertical"/ "horizontal") to individual texture exemplars presented in blocks of 96 trials. Target orientation was random and balanced across blocks. All texture stimuli were presented briefly at maximum contrast, and were followed by a matched pattern mask and feedback tone. Presentation duration was varied across blocks to obtain psychometric functions.

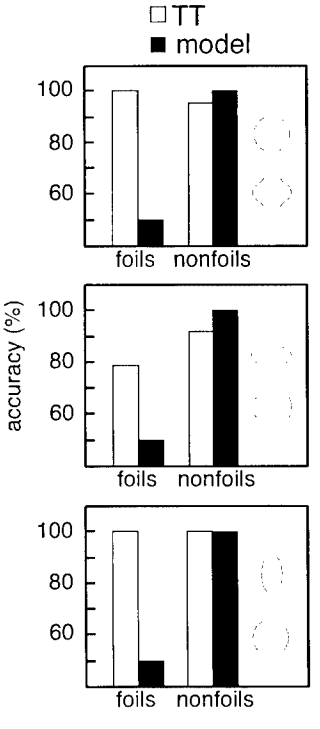

Figure 5. Segregation accuracy for three texture stimuli, for a human observer and an optimal generalized channel energy model. Stimulus duration was 200 ms.

3.2 Results

Figure 4 shows the texture segregation performance for two subjects on three different textures. The shapes of the elements in the textures are shown on the right. The results demonstrate that even for brief presentations the human observers are performing better on these stimuli than the optimal channel energy model.

However, it is possible that during the course of the experiment subjects are learning and using histogram differences for the specific stimuli in the experiment. To test this possibility, we created stimulus ensembles that consisted of approximately 10% foils and 90% non-foils. We computed the optimal histogram model for these particular stimuli and then determined the performance of the optimal model on these same stimuli (which overestimates of the model's performance). Figure 5 shows the performance of the model and one of the subjects on these ensembles. The subject outperformed the optimal channel histogram model, even though the model was being given an unfair advantage.

3.3 Discussion

This experiment demonstrates that human observers can, in brief presentations, segregate image regions based solely on differences in local image structure. This result is undoubtedly quite general because the texture elements in this experiment were picked arbitrarily. Indeed, we have similar preliminary data for other element shapes.

Although generalized channel energy models represent some information about local image structure (i.e., phase information), they do not represent sufficient structural information to segregate the class of stimuli described here. This creates difficulties for all channel energy models proposed to date, because these models do not segregate as accurately as the optimal generalized channel energy model considered here.

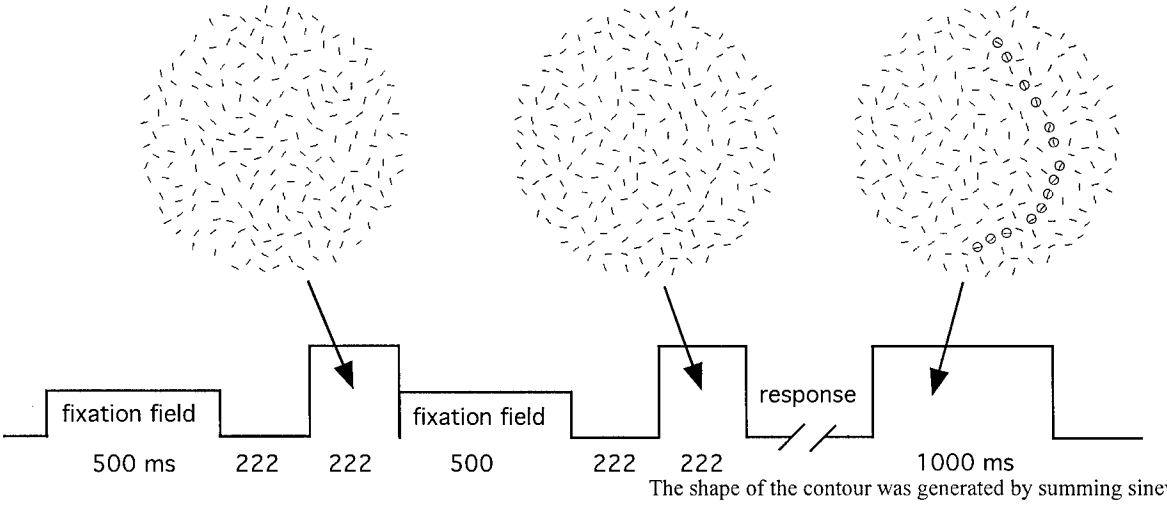

Figure 6. Example stimuli and presentation sequence for the contour detection experiment (Experiment 2).

Of course, it is intuitively obvious that the human visual system extracts precise descriptions of local spatial structure. What the experiments here demonstrate is that the visual system uses such descriptions to perform fast ("preattentive") region segregation. Although not demonstrated here, image structure models can segregate the kinds of textures considered in this experiment.

4. EXPERIMENT 2

The second major class of tasks that pose a difficulty for channel energy models are those that involve grouping of regions containing smoothly changing image structure, such as smooth contours. To obtain additional systematic data on human ability to group contour information, and to provide a test of the image structure model, we measured contour detection performance as a function of contour shape and length. This is a particularly useful task because it involves complex naturalistic judgements under high degrees of uncertainty; yet, the predictions depend upon very few parameters in the image structure model.

4.1 Methods

Accuracy was measured in a two interval forced choice task for detection of line-segment contours in a background consisting of randomly oriented line segments. Four properties of the randomly shaped contours were parametrically varied: amplitude, fractal exponent, length, and level of orientation jitter of the contour elements. This family of contours was selected to be representative of a broad range of naturalistic contours.

4.1.1 Stimuli

Figure 6 illustrates the time line for a single trial, including examples of a background and a background + target. The circular display was 12.3 deg in diameter (32.5 pixels/deg), at the viewing distance of 112 cm. The line-segment elements were 0.31 degrees in length. The target contour and background texture were created in a fashion similar to that in Experiment 1. For purposes of creating the stimuli, each line element was represented by a virtual circle with a diameter equal to twice the element length. The virtual circles were randomly placed one at a time in the image with a restriction that if a new virtual circle intersected an existing one then a new random position was selected.

The shape of the contour was generated by summing sinewave components with frequencies that were harmonics of 0.5 cycles per image. The sinewave components always modulated about an axis through the center of the display; the orientation of the axis was random on each trial. Different random contour shapes were obtained for each trial by randomly selecting the amplitudes and phases of the components, and then filtering (i.e., multiplying the amplitudes by a transfer function). The line elements were randomly placed on the contour first. Then the background line elements were added such that the density of line elements in the background was the same as along contour.

Contour detection accuracy was measured parametrically as a function of four variables: (a) the fractal exponent of the amplitude transfer function (1, 1.5, 2, and 3), (b) the RMS amplitude of the contour modulation (6.5%, 12.5%, 25%, and 50% of the display diameter), (c) the contour axis length (20%, 40%, 60% and 80% of the display diameter), and the range of orientation jitter of the elements (0, 30%, 50%, and 70% of the maximum value, 180°).

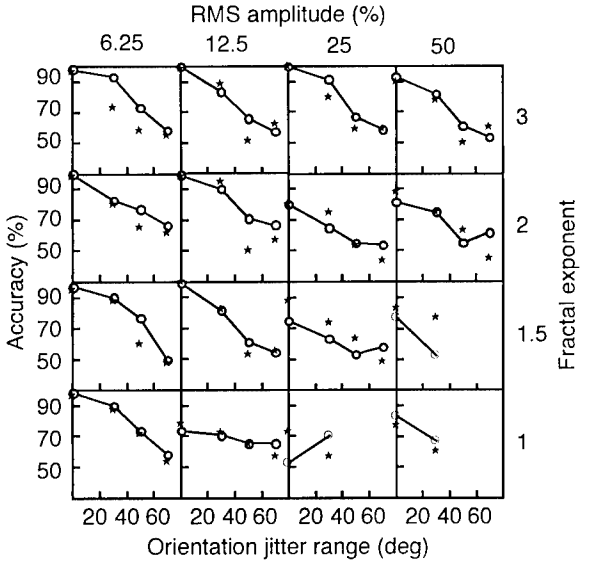

Figure 7. Length = 80% of display diameter. Open circles: contour detection accuracy for random contours, as a function of contour shape (fractal exponent), average contour amplitude (RMS amplitude), and magnitude of orientation jitter of the elements (Orientation jitter range). Solid stars: predicted performance of a two-parameter image structure model.

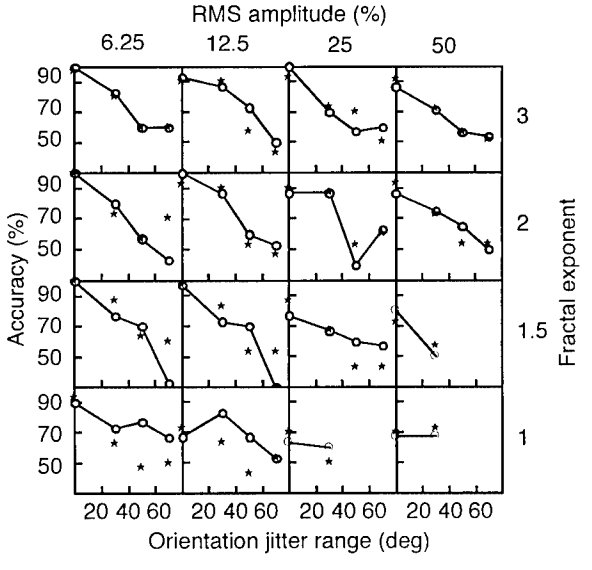

Figure 8. Length = 60% of display diameter. See Figure 7.

4.1.2 Procedure

As shown in Figure 6, on each trial the fixation cross was extinguished 222 ms before presentation of the two test intervals, which were each 222 ms in duration and separated by 720 ms. After the subject responded, he was informed about the correctness of the response, and shown the actual location of the contour.

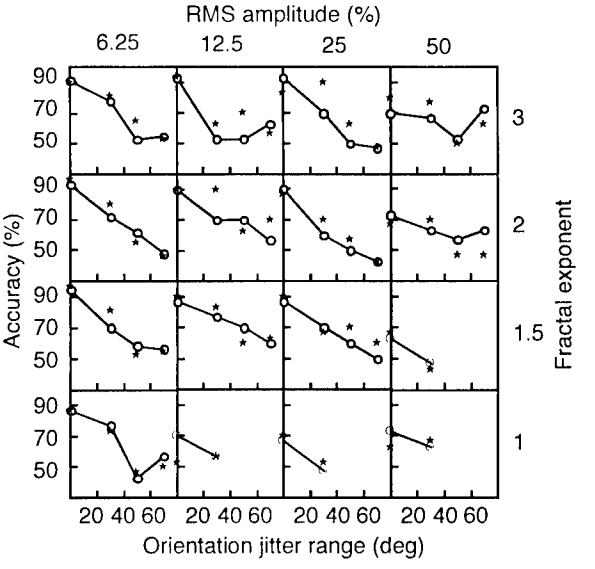

Figure 9. Length = 40% of display diameter. See Figure 7.

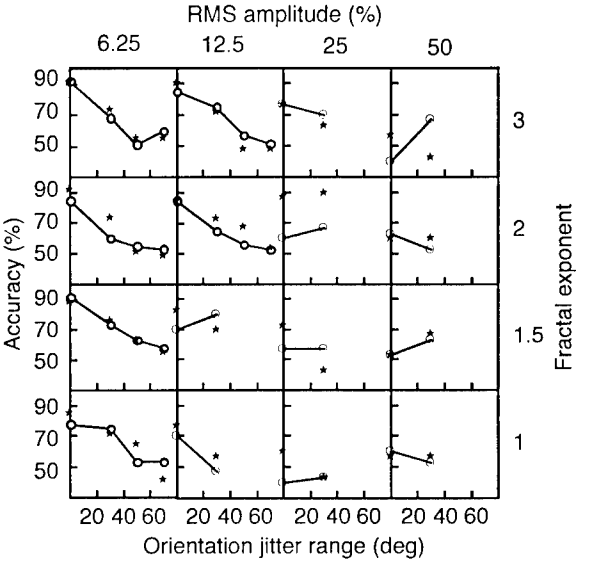

Figure 10. Length = 20% of display diameter. See Figure 7.

Each experimental session consisted of 16 blocks of 30 trials and lasted approximately 40 min. In each block, the stimulus parameters were held fixed. The order of conditions was picked to minimize systematic practice and fatigue effects. All conditions were repeated twice for a total of 60 trials per condition per subject.

4.2 Results

Figures 7-10 show the results for two subjects. Each figure is for a different contour length. The open circles in each panel within a figure show the average accuracy for the two subjects as a function of the range of element orientation jitter. (At the time of writing, the data for this experiment were not complete, so some data points represent results for only one subject.) The solid stars show the predictions of an image structure model described below. Across rows, the plots are for contours of different fractal exponents. Across columns the plots are for contours of different RMS amplitudes. The panels with only two points indicate conditions where all non-zero levels of jitter were not tested, based

upon pilot experiments showing that performance was poor even with 0% jitter.

There are some obvious trends in the data. Performance generally improves with increases in the fractal exponent and contour length, and generally declines with increases in RMS amplitude and jitter. Although these trends are not surprising, the specific levels of performance in the different conditions provide strong constraints on models of contour detection.

4.3 Discussion

This experiment provides a parametric overview of human capabilities for detecting contours in noisy backgrounds. As the data show, humans are quite good at detecting contours even when there is great uncertainty in location, orientation, and shape of the target contour. For example, contours like that in the middle picture of Figure 6 are detected with better than 90% accuracy.

It is generally acknowledged both by perception and computer vision researchers that humans have a remarkable ability to detect spatial structure and statistical regularities in images, even when the structure is unfamiliar (e.g., see Witkin and Tannenbaum, 1982). Thus, our initial assumption was that the relatively simple class of image-structure models described here would be unable to achieve the performance levels of humans in the present experiment. Our aim in testing the models was to identify conditions where the models fail to predict performance, with the hope that these failures might provide hints about what additional mechanisms the human visual system may be using. To our surprise, a simple image structure model (with just two free parameters) does a good job of accounting for the data, suggesting that human contour detection performance in these types of displays may be largely explained by relatively simple, essentially "bottom-up," processing.

In the specific model described here, the input was taken to be all the individual line segments in the display (not the individual pixels). This is equivalent to assuming that primitive detection and initial grouping have already found the groups corresponding to the individual line segments. Thus, the primary computations in the model consisted of a matching stage which compared line segments and a subsequent grouping stage which bound them into groups (see Figure 2).

Because the line segments were all of the same size and form, equation (1) reduces to the weighted sum of the position difference, orientation difference and continuation difference. Furthermore, we found that the information extracted by continuation difference (which had not been considered in Geisler & Super, 1996/99) was redundant with the orientation difference, so we could eliminate the orientation difference. Thus, equation (1) reduces to:

$$D = w_p D_p + w_c D_c \qquad (5)$$

The value of D was computed for each possible pairing of line segments in the image. Groups were then formed by applying associative grouping for a particular value of the binding criterion, β. In other words, we computed D_{ij} for each possible pairing of line-segment groups, g_i and g_j, and then applied the associative grouping rules given by equations (3) and (4). This was done for the stimuli in both intervals in the forced choice presentation. The longest group obtained in each interval was then selected. Which ever of these two groups had more elements was picked as the interval containing the contour. If the two groups happened to have the same number of elements then the group with the smallest summed grouping differences (the strongest binding) was picked as the interval containing the contour.

There are only two parameters in this model: one of the dimensions weights, w_p, and the binding criterion, β. The other dimension difference weight, w_c, is not free because the weights sum to 1.0.

We now define the group difference measures. These particular measures were devised on the basis of intuition, and a little trial and error. No great significance should be attached to these specific formulas. Undoubtedly, there are other related formulas which would capture the same information about the differences between line segments.

The position difference, D_p, was taken to be the Euclidean distance between the nearest pixels in the two line segments. We used this measure based upon experimental results of Geisler & Super (1996/99), who found that proximity grouping is better described by the near point distance than by the distance between object centroids. Note that because of the position difference component, the computation of the total difference, D, is a local process; only neighboring line segments influence the groups that are formed.

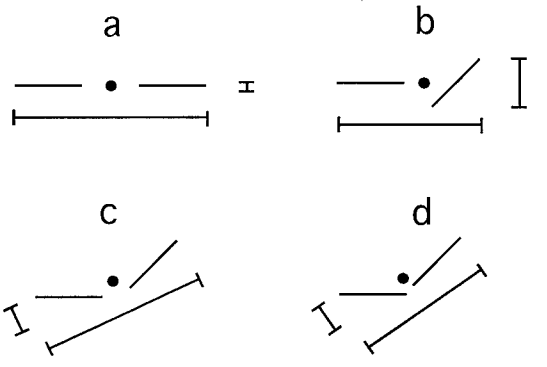

Figure 11. Illustration of the continuation difference measure. Each sub-figure shows a pair of line segments in some configuration. The solid dot shows the centroid of the group formed by the two segments. The long "error bar" shows the orientation of the major axis; the short "error bar" the orientation of the minor axis. The position difference is the ratio of the standard deviations computed along the two axes (through the centroid); this quantity is approximately the length of the short "error bar" divided by the length of the long "error bar."

The continuation difference, D_c, is a straight-forward measure meant to capture (in conjunction with the position difference measure) the degree to which two line segments could be smooth continuations of each other. Note first that any two line segments can be considered as a single group of points. This group will have a centroid, and a major axis, which is the best fitting line passing through the centroid. The continuation difference is defined to be the ratio of the standard deviation of the distance of the points from the major axis divided by the standard deviation of the distance of the points from the minor axis (the axis perpendicular to the major axis):

$$D_c = \frac{\sigma_M}{\sigma_m} \qquad (6)$$

Note that this measure varies between 0 and 1. Figure 11 illustrates how this measure behaves. The longer "error bar" in each sub-figure indicates the orientation of the major axis, the shorter "error bar" the orientation of the minor axis, and the solid dot the centroid. The continuation difference is roughly proportional to the ratio of the length of the short "error bar" to the long "error bar." When the two line segments fall along a straight line (Figure 11a) then the standard deviation about the major axis is zero, and hence the continuation difference is zero. If the second line segment is rotated (Figure 11b) the standard deviation about the major axis increases and so does the continuation difference. If the second line keeps the same orientation difference but is shifted vertically so that it is more consistent with a smooth contour (Figure 11c), then the continuation difference decreases. For a given orientation difference between the line segments, the closer the line segments the greater the continuation difference (compare Figures 11c & 11d). This is consistent with the fact that a greater curvature would be required to connect the two line segments when they are closer.

The model contained one additional constraint. Although the contours were random in orientation, shape, and position on every trial in a block, there was always some limitation on the possible locations of the contours. These limitations were obvious to the subjects when running the experiment. The model was also given this information; it did not consider groups which fell outside the region of possible contour locations.

To estimate the best fitting parameter values, a coarse grid search was followed by a more refined grid search at the most promising locations. For each pair of parameter values, the performance of the model was computed for exactly the same stimuli that the subjects saw.

The solid stars in Figures 7-10 show the predictions of the model. As can be seen, the model does a remarkably good job of predicting the performance across all the conditions. The estimated parameter values are as follows: $w_p = 0.2$, $w_c = 0.8$, $\beta = 0.17$.

Importantly, these parameter values, which fit the human data best, are also the values that maximize the absolute accuracy of the model in the task. In other words, combining local measurements of distance and continuation so that they produce the greatest accuracy in the model, also yields a model performance that is close to human performance. This fact adds some support for this class of models.

Further, the results suggest that human contour detection in these types of displays may involve only local measures of group differences followed by an unsophisticated grouping mechanism, such as associative grouping. Something similar to the local grouping difference measures and associative grouping should be relatively easy to implement neurally.

Finally, if the human visual system is using these simple mechanisms then the results imply that it has evolved or learned nearly optimal weights for combining local difference information and nearly optimal criteria for controlling associative grouping.

5. CONCLUSION

The experiments and analyses reported here demonstrate that there are many texture grouping and segregation situations that are difficult to model within the framework of the generalized channel energy models. The heart of the difficulty for such models is that the human visual system extracts detailed local spatial structure and is able to use it for grouping and segregation. Unfortunately, we see no way to avoid the difficult problem of modeling how the visual system extracts and represents spatial structure. Although image structure models cannot yet be easily applied to arbitrary images, we have made a start, and we have demonstrated that they can be quite effective in limited domains. In particular, a very simple image structure model that combines local measures of proximity and continuation can account for human ability to detect random contours that are representative, in complexity and uncertainty, of those occurring in the natural environment.

6. REFERENCES

Bergen, J. R. "Theories of visual texture perception." In D. Regan (Ed.), *Vision and visual dysfunction* (Vol. 10B: Spatial vision, pp. 114-134). New York: Macmillan, 1991.

Bovik, A. C., Clark, M., and Geisler, W. S. "Multichannel texture analysis using localized spatial filters." *Pattern Analysis and Machine Intelligence, 12*, 55-73, 1990.

Estabrook, G.F. "A mathematical model in graph theory for biological classification." *Journal of Theoretical Biology*, 12, 297-310, 1966.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Research, 33*(2), 173-193.

Geisler, W.S. and Super, B.J. Perceptual organization of two-dimensional patterns. *Psychological Review*, under review.

Geisler, W.S. and Super, B.J. Perceptual organization of two-dimensional patterns. UT-CVIS-TR-96-002, Austin, Texas: Center for Vision and Image Sciences, 1996.

Graham, N., Beck, J., and Sutter, A. "Nonlinear processes in spatial-frequency channel models of perceived texture segregation: effects of sign and amount of contrast." *Vision Research, 32*(4), 719-743, 1992.

Sha'ashua, S., and Ullman, S. *Structural saliency: The detection of globally salient structures using a locally connected network.* Paper presented at the Proceedings of the Second International Conference on Computer Vision, 1988.

Thornton, T. and Geisler, W.S. "Texture segregation on the basis of local shape information." *Investigative Ophthalmology & Visual Science Supplement*. (ARVO) 39/4, S649, 1998.

Victor, J.D. and Brodie, S.E. "Discriminable textures with identical Buffon needle statistics." Biological Cybernetics, 31, 231-234, 1978.

Witkin, A. P., & Tenenbaum, J. M. On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and Machine Vision* (pp. 481-543). New York: Academic Press, 1983.

7. ACKNOWLEDGEMENTS

This work was supported by grants from the National Eye Institute, NIH.